DOCUMENT RESUME

ED 281 882 TM 870 289

AUTHOR Miao, Chang Yu
TITLE A Bayes' Subjective Probability Approach to the Nedelsky Procedure.
PUB DATE Apr 87
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; Analysis of Variance; *Bayesian Statistics; *Cutting Scores; Higher Education; Item Analysis; Measurement Techniques; Minimum Competencies; *Multiple Choice Tests; Reliability; Response Style (Tests); *Test Theory
IDENTIFIERS *Nedelsky Method; Standard Setting; *Subjective Probability

ABSTRACT

Nedelsky (1954) has suggested a procedure for determining the minimum passing score on a multiple-choice test. In this procedure expert judges estimate the probable score of a minimally competent examinee. The technique does not refer to the students' performance data. The purposes of this paper are: (1) to introduce a modification to the Nedelsky procedure that includes both absolute and relative standard setting procedures through the use of subjective probabilities; and (2) to compare these two procedures. The modified Nedelsky method (i.e., the Bayes Subjective Probability method) uses the students' response patterns on each test item and estimates the minimal item value by taking both the students' performance and the experts' judgments into consideration. The Nedelsky procedure has only discrete probabilities ranging from 0 to 1.0; however, the Bayes Subjective Probability procedure has a continuous distribution ranging from 0 to 1.0. Analysis of variance based reliabilities are both high (0.96 for Nedelsky and 0.94 for Bayes Subjective Probability method). For the Nedelsky method, a higher cutting score value emerges and a higher variability index is found. Data for this study were collected from 186 Indiana University students who took the Spanish version of the Foreign Language Placement Examination and from seven Spanish Department faculty members who judged the test items. Examples of comparative data for five test items are appended. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281882

A Bayes' Subjective Probability Approach to the Nedelsky Procedure

Chang Yu Miao

Bureau of Evaluative Studies and Testing

Indiana University

Bloomington

Prepared for the annual meeting of the

National Council on Measurement in Education

Washington D. C.

1987

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Chang Yu Miao

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

TM 870 289

A Bayes' Subjective Probability Approach to the Nedelsky Procedure

Abstract

Nedelsky (1954) has suggested a procedure for determining the minimum passing score on a multiple-choice test. In this procedure expert judges estimate the probable score of a minimally competent examinee. The technique does not refer to the students' performance data. The purposes of this paper are 1) to introduce a modification to the Nedelsky procedure that includes both absolute and relative standard setting procedures through the use of subjective probabilities, and 2) to compare these two procedures. The modified Nedelsky method (i.e.,the Bayes Subjective Probability method) uses the students' response patterns on each test item and estimates the minimal item value by taking both the students' performance and the experts' judgments into consideration. The Nedelsky procedure has only discrete probabilities ranging from 0 to 1.0; hovever , the Bayes Subjective Probability procedure has a continuous distribution ranging from 0 to 1.0 . Analysis of variance based reliabilities are both high (0.96 for Nedelsky and 0.94 for Bayes Subjective Probability method). For the Nedelsky method, a higher cutting score value emerges and a higher variability index is found.

A Bayes' Subjective Probability Approach to the Nedelsky Procedure

Standard setting, or setting a passing score, poses a considerable psychometric problem . Traditionally, the decision was made through a normative approach that set a passing score at a certain distance from the mean of total scores in standard deviation units. The minimum score actually needed for passing the test is fixed at one or two standard deviations below the mean, regardless of the considerations of how the test items were composed and how the competencies of examinees are to be distributed.

For the absolute standard setting approach using multiple-choice items, an absolute standard is needed to evaluate students' learning. Various methods have been developed (Nedelsky, 1954; Angoff, 1971; Ebel, 1979). The existing models which use a group of judges to classify or examine item content originated from the method presented by Nedelsky in 1954. According to the Nedelsky procedure, judges reject the incorrect choices of a multiple-choice test item that a borderline student should be able to eliminate. After all judges reach agreement, an expected minimal passing score can be set up by adding the reciprocal of the number of the options not rejected by the judges for each item. There is considerable research dealing with the comparisons among these procedures (Skakun and Kling, 1980; Brennan and Lockwood, 1980; Saunders, Ryan and Huynh, 1981).

Glass (1978) identified six classes of methods for reaching a compromise cutting score on a criterion-reference test. Beuk (1984) used a simple linear function to relate the absolute passing score and the relative passing score. A procedure suggested by Jaeger (Cross, Impara, Frary and Jaeger, 1984) required that normative data be provided to the judges while evaluating the items. Only the judges' judgments were used

judges while evaluating the items. Only the judges' judgments were used by these researchers although attempts were made to include normative information in setting standards. None of them made direct use of students' response patterns when setting cutting scores.

When the judges are asked the question: " Should the minimally competent students be able to identify this option as correct ?", the judges always have a reference population in their mind. Therefore, there seems to be rationale to include the students' response patterns in setting cutting scores. The purpose of this study was to combine the absolute and relative information in defining cutting scores by considering both the difficulty of the examination and the response patterns of the examinees. In this paper, using experts' judgments and students' performance , a conditional probability can be found. Then through Bayes' rule, the borderline student's performance on each item is estimated. Statistics of both the Nedelsky procedure and the subjective probability modification are compared.

Nedelsky Procedure

The conventional letter grades F, D, C, B and A are used to define minimal competency in the Nedelsky procedure with the meaning of failure (F), barely passing (D). The procedure requires judgments by experts considering the performance of hypothetical barely passing students on each item of a test. Before the test is administered, experts are given copies of the test and are asked to cross out the responses which the lowest D-student should be able to reject as incorrect. To the left of the item , the reciprocal of the number of the remaining responses is recorded. After the teachers have marked all the test items, they hold a

brief conference to compare and discuss discrepant decisions. After such a conference, the sum (M) of those reciprocals of the numbers of responses are computed for each judge. The mean ($\overline{M}$) and standard deviation (S) are determined. Then the score corresponding to the borderline student is set equal to $\overline{M} + k * S$, where k is a constant whose value is determined by the judges.

Bayes' Subjective Probability Approach

The proposed subjective probability modification is based on two assumptions:

1. The borderline students will choose randomly among the alternatives selected by each judge .
2. Each judge of the standard setting group has the same opinion about what should be required for a borderline student.

The consequence of the first assumption is to estimate the probability of borderline students choosing the correct answer. The second assumption ensures that all the judges have the same opinion about the requirements for borderline students .

In describing the method of estimating the cutting score, the following definitions are made :

a. P1 : probability of answering the item correctly by a borderline student.

b. P2 : probability of a borderline student among those students who answered the item correctly.

c. P3 : probability of a borderline student among those students who answered the item incorrectly.

d. P : probability of answering the item correctly by all the students.

e. Q : probability of answering the item incorrectly by all the students. Q = 1-P

According to Bayes' rule, the relationship among these five probabilities is :

$$P1 = \frac{P2 * P}{P2 * P + P3 * Q}$$

In order to calculate P1, it is necessary to determine values for P, Q, P2 and P3. Values of P and Q are easily found by using the item difficulty index. That is, P is the proportion of all students who answer the item correctly , and Q is 1 - P.

The calculation of P2 is found by the formula :

$$P2 = \frac{\text{the expected number of borderline students who answer correctly}}{\text{total students who answer the correct option}}$$

*note: 'the expected number of borderline students who answer correctly' is obtained by

$$\frac{\text{total students} - \text{the number of students who answer the incorrect alternatives eliminated by a judge}}{\text{the number of alternatives selected by a judge}}$$

If the 'expected number of borderline students who answer correctly' is more than actual number of students who answer correctly , then the probability of a borderline student who answer this question correctly is

set to be 0 .

The computation of P3 is found by the formula :

$$P3 = \frac{\text{total number of students who choose incorrect alternatives identified by a judge}}{\text{(total students)- (students who answer correctly)}}$$

Illustration:

Suppose an item is being judged to set up an expected score for a borderline student. Here is an example based on the judgments of a single judge from the pool of judges. Alternative B is eliminated by the judge. The number of students who answered each alternative is listed below that alternative.

alternatives	A	B	C	D*
number of students	4	22	12	104
judgment	1	0	1	1

total students = 186
(44 students left blank)

Legend:

a. judgment 0 is for the alternatives rejected by a judge. i.e., the judge believed minimally qualified students would see it as wrong.
b. judgment 1 is for the alternatives not rejected by a judge
c. * stands for correct answer

Calculation:

P = 104/186= 0.56
Q = 1 - 0.56 = 0.44
P2 = ((186 - 22) * (1/3)) / 104 = 0.52
P3 = (4 + 12) / (186-104) = 0.20
P1 = (0.52*0.56) / ((0.52*0.56)+(0.20*0.44) = 0.76

The probability of a borderline student answering this item correctly

based on this expert judge and this set of students' response is .76 . In contrast, the probability of a borderline student answering this item correctly based only on this expert judge is .33 using the Nedelsky method.

---

alternatives	A*	B	C	D
numbers of students	17	46	19	30
judgement	1	1	1	0

total students = 186
(74 students left blank)

Calculation:

The expected number of borderline students answering the question correctly is :

(186-30) * (1/3) = 52 > 17

therefore, the probability of borderline students answering this question correctly is set to be 0.

Method

A. Data Source

The data for this study were collected from students who took the Spanish version of the Foreign Language Placement Examination administered by the Bureau of Evaluative Studies and Testing (B.E.S.T.) at Indiana University. The test published by Educational Testing Service has been used for this study. The test has a reading and listening subtest but only the reading part has been utilized. The examination had 60 items and had four alternatives for each item.

four alternatives for each item.

B. Judges

Seven teachers in the Spanish Department were used for this study. All of the teachers spoke Spanish at the native level and each had several years of teaching experience. The group consisted of three males and four females.

C. Subjects

The 186 subjects were the students who had taken prior work (most in high school) in Spanish. The sample included all the students registered for the placement examination in Spanish courses during 1986 summer session.

D. Procedure

All 186 students took the Spanish placement examination under the supervision of test administrators. Before the test was given, the seven Spanish teachers were scheduled for the Nedelsky procedure. Each teacher worked independently after reading the rules for making judgments (Nedelsky, 1954). Together with the students' data, those judgments were analyzed using a computer program of the Bayes' formula written by the author.

E. Analysis

The data were analyzed by the SPSS package and the author's program. Item total score distribution was obtained. For the Nedelsky and the Subjective Probability methods the probability of borderline students answering the item correctly was calculated for each item. Then the

cutting score of each judge, and the average of those cutting scores across all judges ,was calculated. The reliabilities of the probability given to each item by both methods were calculated by the formula presented by Winer (1962), and used by Skakun and Kling (Skakun and Kling,1980).

$$r_k = 1 - \frac{\text{MS within items}}{\text{MS between items}}$$

F. Research Questions

The analyses were designed to respond to four questions .

1. What are the cutting scores with respect to the Nedelsky and the Bayes Subjective Probability methods ?

2. What are the reliabilities of the probability of answering the item correctly by borderline students for the Nedelsky and the Bayes Subjective Probability methods ?

3. What are the variances of cutting scores among these judges for the Nedelsky and the Bayes Subjective Probability methods ?

4. Are there differences between cutting scores of these methods ?

Results

From Table 1, the students' performances were tested with chi-square goodness-of-fit test (alpha = .05) and appear to be normally distributed.

table 1

The Frequency Distribution

raw score	frequencies	cumulative frequencies
45	1	186
27	1	185
25	2	184
24	2	182
23	3	180
22	1	177
21	3	176
20	4	173
19	4	169
18	8	165
17	9	157
16	16	148
15	11	132
14	12	121
13	14	109
12	12	95
11	18	83
10	17	65
9	13	48
8	11	35
7	9	24
6	6	15
5	2	9
4	5	7
3	1	2
0	1	1

total number of items = 60
mean of percentage right score = 12.88
standard deviation of percentage right score = 5.58

Table 2 and Table 3 show the results by the Nedelsky procedure and by the subjective probability approach.

Table 2

The Cutting Score Set by Each Judge

judge	by the Nedelsky procedure	by the Subjective Probability procedure
1	15.58	7.65
2	19.58	9.17
3	24.16	9.35
4	36.66	27.01
5	39.83	26.66
6	22.16	8.08
7	25.40	9,89
mean of cutting score	26.19	13.97

Table 3

Descriptive Statistics

	Nedelsky	Subjective Probability
range of cutting score	15.58 - 39.83	7.65 - 27.01
mean of cutting socre	26.19	13.97
standard deviation	8.88	8.81
reliability	0.969	0.943

total number of items = 60
total number of students = 186
total number of judges = 7

The cutting scores and the percentages of passing and failing associated with each method were presented.

Table 4

Percentage of Passing and Failing

	Nedelsky	Subjective
passing score	26.19	13.97
% passing	2 %	42 %
% failing	98 %	58 %

Discussion

Based on the t-test, results indicated that there was a significant difference of the cutting scores between the cutting scores of the Bayes Subjective Probability method and the Nedelsky method (alpha=.05). Although it is not possible to know the actual percentage of border-line students among these examinees, the failure rate (98%) according to the Nedelsky method seems unrealistically high. The Bayes Subjective Probability method had a much lower failure rate (58%) than the Nedelsky method and was much closer to the failure rate (51%) set by the Spanish department (i.e., at the mean). In summary, the Bayes Subjective Probability modification is different from the original Nedelsky procedure : 1) in the continuum of the probability value of borderline students answering the item correctly between 0 to 1, 2) in variability ,and 3) in comparison of the cutting score to the standard set by the Spanish Department (12.88). These features give credence to the use of the Bayes Subjective Probability Approach.

When there are relatively few students answering an item correctly (i.e., an item more difficult than chance), the Bayes Subjective Probability method gives the probability of the borderline student answering that item a value 0 ,but the Nedelsky gives a value no less than .25. For this test administration there were many items that fell

than .25. For this test administration there were many items that fell into this category. Because of the small number of judges involved in this study, further investigation is needed. The index used to select judges developed by van der Linden (1982) could also be used together for further investigation.

REFERENCE

Angoff, W.H., Scales, Norms, and Equivalent Scores. In R.L. Thorndike, Eduational Measurement. 1954, Washington, D.C. American Concil on Educatiuon.

Beuk, C.H., A Method for Reaching A Compromise Between Absolute And Relative Standards In Examinations. Journal of Educational Measurement, 1984, v21, n2, 147-152.

Brennan, R.L., & Lockwood, R.E. A Comparison of the Nedelsky and Angoff Cutting Score Procedure Using Generalization Theory. Applied Psychological Measurement, 1980, 4, 219-240.

Cross, L.H., Impara, J.C., Frary, R.B. & Jaeger, R.M. A Comparison of Three Methods for Establishing Minimum Standards on the National Teacher Examinations. Journal of Education Measurement, 1984, v21, n2, 113-129.

Ebel, R.L., Essentials of Educational Measurement. Englevood Cliff N.J.; Prentice-Hall, 1972.

Glass, G.V., Standards and Criteria, Journal of Educational Measurement, 1978, v15, n4, 237-261.

Nedelsky, L. Absolute Grading Standards for Objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Sauders, J.C., Ryan, J.P. and Huynh, H. A Comparison of Two Approaches to Setting Passing Scores Based on the Nedelsky procedure. Applied Psychological Measurement, 1981, v5, n2, 209-217.

Skakun, E.N., Kling S. Comparability of Method for Setting Standards. Journal of Educational Measurement, 1980, 17, 229-235.

van der Linden, W.J. A Latent Trait Method for Determining Intrajudge Inconsistency in the Angoff and Nedelsky Techniqus of Standard Setting. Journal of Educational Measurement, 1982, v19, n4, p295-308.

Winer, B.J. (1962) Statistical Principles in Experimental Design. New york, McGrav-Hill Book Company.

Appendix

Legend for the following examples

'0' stands for the alternatives rejected by a judge
'1' stands for the alternatives not rejected by a judge
'*' stands for the correct answer
'p' stands for item difficulty
'q' is 1-p
'N' total number of students

Example 1

alternatives	A	B*	C	D	Bayes Subjective	Nedelsky
number of students	35	0	35	30		
judgment	0	1	0	0	0.	1.00
judgment	1	1	0	0	0.	0.50
judgment	1	1	1	0	0.	0.33
judgment	1	1	1	1	0.	0.25

p = 0.
q = 1.0
N = 100

Example 2

alternatives	A	B*	C	D	Bayes Subjective	Nedelsky
number of students	25	25	25	25		
judgment	0	1	0	0	1.00	1.00
judgment	1	1	0	0	0.50	0.50
judgment	1	1	1	0	0.33	0.33
judgment	1	1	1	1	0.25	0.25

p = 0.25
q = 0.75
N = 100

Example 3

alternatives	A	B*	C	D	Bayes Subjective	Nedelsky
number of students	15	50	15	20		
judgment	0	1	0	0	1.00	1.00
judgment	1	1	0	0	0.69	0.50
judgment	1	1	1	0	0.47	0.33
judgment	1	1	1	1	0.33	0.25

p = 0.50
q = 0.50
N = 100

Example 4

alternatives	A	B*	C	D	Bayes Subjective	Nedelsky
number of students	8	75	8	7		
judgment	0	1	0	0	1.00	1.00
judgment	1	1	0	0	0.84	0.50
judgment	1	1	1	0	0.67	0.33
judgment	1	1	1	1	0.57	0.25

p = 0.75
q = 0.25
N = 100

Example 5

alternatives	A	B*	C	D	Bayes Subjective	Nedelsky
number of students	0	100	0	0		
judgment	0	1	0	0	1.00	1.00
judgment	1	1	0	0	1.00	0.50
judgment	1	1	1	0	1.00	0.33
judgment	1	1	1	1	1.00	0.25

p = 1.00
q = 0.
N = 100

Figure 1 p increased for a specific pattern of judgment

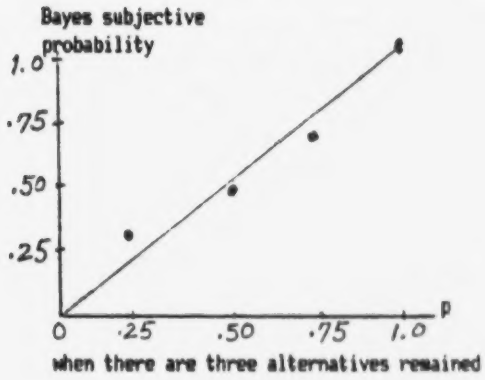

Figure 2 alternatives rejected by judge decreased for a specific p value